Parcial 2 – Procesamiento de datos a gran escala

Brayan Steven Carrillo Mora
Santiago Rueda Pineda.

Procesamiento de datos a gran escala

Programa de Ingeniería de Sistemas

Facultad de Ingeniería

Pontificia Universidad
JAVERIANA
Colombia

Bogotá D.C, Colombia

Mayo 2024

Tabla de Contenido

1. Objetivo

El objetivo de este proyecto es aplicar y poner en práctica los conceptos aprendidos en la asignatura de Procesamiento de Datos a Gran Escala, utilizando técnicas de machine learning para analizar y responder preguntas específicas basadas en dos conjuntos de datos reales sobre nacimientos en los municipios de Acacías y Buga, Colombia. A través de este análisis, se pretende demostrar la capacidad de procesamiento y extracción de información relevante de grandes volúmenes de datos, facilitando la toma de decisiones informadas y el entendimiento profundo de las tendencias y patrones demográficos en estas regiones.

2. Contexto

El equipo (máximo 3 personas) tiene que desarrollar un estudio analítico sobre los nacimientos en Colombia. Existen en la actualidad registros de nacimientos en diferentes municipios de Colombia, con diferentes características almacenados en el portal www.datos.gov.co .La idea es seleccionar y unir al menos 2 municipios diferentes (2 conjuntos de datos diferentes), y elaborar su proyecto en función de una idea para aplicar dos técnicas de Machine Learning de su preferencia.

3. Descripción

En este capítulo, se busca describir el proceso realizado durante el ejercicio del parcial, mostrando las diferentes fases de abordaje de los integrantes del grupo, además de los detalles necesarios para elaborar cada ítem.

a. Fase Inicial

- **Selección de los conjuntos de datos y análisis de su contexto**

El primer paso del proyecto fue seleccionar los conjuntos de datos de nacimientos de dos municipios diferentes. La selección se basó en la disponibilidad y calidad de los datos, así

como en la diversidad de características que permitan un análisis enriquecedor. En este caso, los conjuntos de datos seleccionados son:

- Nacidos vivos en el Municipio de Acacías
- Guadalajara de Buga: Nacidos Vivos 2021

La elección de estos conjuntos de datos, cuyos enlaces pueden ser encontrados en las referencias, fue dada por la relativa similitud en sus variables, además de las características de ambos municipios, los cuales gozan de condiciones climáticas similares, una densidad poblacional similar, además que ambas representan municipios en crecimiento en sus respectivos departamentos.

Se realizo un análisis inicial del contexto de estos datos para identificar cómo pueden complementarse y qué tipo de información única aporta cada conjunto de datos.

- **Identificación de Preguntas de Investigación**

El equipo propuso las siguientes dos preguntas de investigación para ser respondidas mediante el análisis de los datos y la aplicación de técnicas de Machine Learning. Estas preguntas deben ser importantes tanto desde un punto de vista práctico como académico, contribuyendo a resolver un problema existente o mejorando un proceso actual.

- Preguntas para resolver

1. ¿Es posible predecir el tipo de nacimiento (parto normal, cesárea, etc.) de un recién nacido en los municipios de Acacías y Guadalajara de Buga utilizando un modelo de clasificación?
2. ¿Se puede predecir, a través de un modelo de clasificación, cuándo un recién nacido forma parte de un nacimiento múltiple en estos municipios?

- **Preparación de los datos**

Para el ejercicio se llevó a cabo una preparación exhaustiva de los conjuntos de datos proporcionados, correspondientes a los nacimientos en los municipios de Acacías y Buga, Colombia. La preparación de datos es una etapa crucial en el procesamiento de datos a gran escala, ya que permite obtener una comprensión profunda de las características y estructuras presentes en los datos. A través de técnicas de análisis descriptivo y visualización, se identificarán patrones, tendencias y posibles anomalías que proporcionarán una base sólida para las etapas posteriores del proyecto. Este análisis preliminar también servirá para detectar y abordar problemas de calidad de los datos, como valores faltantes o atípicos, garantizando así la fiabilidad y precisión de los resultados obtenidos.

Durante la preparación de datos, se realizaron las siguientes acciones**:**

- Cambio de Nombre de Columnas: Como se puede observar en la Grafica 1 – Muestra de código en el cambio de variables se renombraron las columnas para asegurar consistencia y claridad en los datos. Esto facilita la comprensión y el manejo de la información durante el análisis.

```
# Renombrar columnas de DF_A
MLDFA = MLDFA.withColumnRenamed("area_nacimiento", "departamento_nacimiento") \
             .withColumnRenamed("sitio_nacimiento", "rea_nacimiento") \
             .withColumnRenamed("nombre_instituci_n", "nombre_instituci_n_nacimiento") \
             .withColumnRenamed("sexo", "g_nero") \
             .withColumnRenamed("peso_gramos", "peso_nacimiento") \
             .withColumnRenamed("talla_cent_metros", "talla_nacimiento") \
             .withColumnRenamed("n_mero_consultas_prenatales", "n_mero_de_consultas_prenatales") \
             .withColumnRenamed("tipo_parto", "tipo_de_parto") \
             .withColumnRenamed("multipilicidad_embarazo", "multiplicidad_de_embarazo") \
             .withColumnRenamed("grupo_sanguineo", "grupo_sanguineo") \
             .withColumnRenamed("factor_rh", "factor_rh") \
             .withColumnRenamed("pertenencia_tnica", "cultura_pueblo_o_rasgos_f") \
             .withColumnRenamed("edad_madre", "edad_madre") \
             .withColumnRenamed("estado_conyugal_madre", "estado_conyugal_madre") \
             .withColumnRenamed("nivel_educativo_madre", "nivel_educativo_madre") \
             .withColumnRenamed("ultimo_a_o_aprobado_madre", "ultimo_a_o_aprobado_madre") \
             .withColumnRenamed("pa_s_residencia", "pa_s_residencia") \
             .withColumnRenamed("departamento_residencia", "departamento_residencia") \
             .withColumnRenamed("municipio_residencia", "municipio_residencia")

# Renombrar columnas de DF_B
MLDFB = MLDFB.withColumnRenamed("departamento_nacimiento", "departamento_nacimiento") \
             .withColumnRenamed("rea_nacimiento", "rea_nacimiento") \
             .withColumnRenamed("nombre_instituci_n_nacimiento", "nombre_instituci_n_nacimiento") \
             .withColumnRenamed("g_nero", "g_nero") \
             .withColumnRenamed("peso_nacimiento", "peso_nacimiento") \
             .withColumnRenamed("talla_nacimiento", "talla_nacimiento") \
             .withColumnRenamed("n_mero_de_consultas_prenatales", "n_mero_de_consultas_prenatales") \
             .withColumnRenamed("tipo_de_parto", "tipo_de_parto") \
             .withColumnRenamed("multiplicidad_de_embarazo", "multiplicidad_de_embarazo") \
             .withColumnRenamed("grupo_sanguineo", "grupo_sanguineo") \
             .withColumnRenamed("factor_rh", "factor_rh") \
             .withColumnRenamed("cultura_pueblo_o_rasgos_f", "cultura_pueblo_o_rasgos_f") \
             .withColumnRenamed("edad_madre", "edad_madre") \
             .withColumnRenamed("estado_conyugal_madre", "estado_conyugal_madre") \
             .withColumnRenamed("nivel_educativo_madre", "nivel_educativo_madre") \
             .withColumnRenamed("ultimo_a_o_aprobado_madre", "ultimo_a_o_aprobado_madre") \
             .withColumnRenamed("pa_s_residencia", "pa_s_residencia") \
             .withColumnRenamed("departamento_residencia", "departamento_residencia") \
             .withColumnRenamed("municipio_residencia", "municipio_residencia")
```

Grafica 1 – Muestra de código en el cambio de variables

- Modificación de columnas redundantes o con datos anómalos: Como se observa en la Grafica 2 – Tratamiento de columnas redundantes, dado que solo estamos evaluando en dos municipios, se rellenan los valores en sus respectivos conjuntos de

datos evitando los errores de digitación y valores nulos que fueron ingresados en la base de datos, al momento de recolectar la información.

```
#Modificamos valores redudante

from pyspark.sql.functions import lit

# Reemplazar todos los valores en la columna "rea_nacimiento" por "Acacias"
MLDFA = MLDFA.withColumn("rea_nacimiento", lit("acacias"))

# Reemplazar todos los valores en la columna "departamento_nacimiento" por "Meta"
MLDFA = MLDFA.withColumn("departamento_nacimiento", lit("Meta"))

#Reemplazar valores
MLDFA = MLDFA.withColumn("rea_nacimiento", lit("acacias"))

MLDFB = MLDFB.withColumn("rea_nacimiento", lit("guadalajara de buga"))

# Mostrar el DataFrame resultante
MLDFA.display()
```

Grafica 2 – Tratamiento de columnas redundantes

- Eliminación de Columnas no equivalentes: En la Grafica 3 – Código para eliminar columnas no equivalentes, se observa la eliminación de columnas o variables que no eran equivalentes entre los diferentes conjuntos de datos. Esto fue crucial para poder fusionar los datos de manera coherente y evitar problemas en el análisis posterior.

```
#Eliminacion de columnas diferentes
# Obtener las columnas que son diferentes entre MLDA y MLDB
columns_in_MLDA_not_in_MLDB = set(MLDFA.columns) - set(MLDFB.columns)
columns_in_MLDB_not_in_MLDA = set(MLDFB.columns) - set(MLDFA.columns)

# Eliminar las columnas que son diferentes entre MLDA y MLDB
for col_name in columns_in_MLDA_not_in_MLDB:
    MLDFA = MLDFA.drop(col_name)

for col_name in columns_in_MLDB_not_in_MLDA:
    MLDFB = MLDFB.drop(col_name)
```

Grafica 3 – Código para eliminar columnas no equivalentes

- Unión de conjuntos de datos: Se procedió a unir los conjuntos de datos con el fin de hacer un buen análisis de datos, esto se puede observar en la Grafica 4 – Unión de los conjuntos de datos.

```
# Unir los dos DataFrames
MLDF = MLDFA.union(MLDFB)

# Mostrar el DataFrame resultante
MLDF.display()
```

MLDF: pyspark.sql.dataframe.DataFrame = [departamento_nacimiento: string, rea_nacimiento: string ... 16 more fiel

Table

	departamento_nacimiento	rea_nacimiento	nombre_instituci_n_nacimiento
1	Meta	acacias	500060016901 HOSPITAL MUNICIPAL DE ACACIAS ES
2	Meta	acacias	500060016901 HOSPITAL MUNICIPAL DE ACACIAS ES
3	Meta	acacias	500060016901 HOSPITAL MUNICIPAL DE ACACIAS ES
4	Meta	acacias	500060016901 HOSPITAL MUNICIPAL DE ACACIAS ES
5	Meta	acacias	500060016901 HOSPITAL MUNICIPAL DE ACACIAS ES
6	Meta	acacias	500060016901 HOSPITAL MUNICIPAL DE ACACIAS ES
7	Meta	acacias	500060016901 HOSPITAL MUNICIPAL DE ACACIAS ES

2,000 rows | 0.68 seconds runtime

Grafica 4 – Unión de los conjuntos de datos

- Exploración de nulos: Tras haber unido los conjuntos de datos para un análisis satisfactorio, se procedió a explorar el conjunto de datos resultante en busca de registros o datos nulos, como se muestra en la Grafica 5 – Código de detección de registros nulos en las columnas del conjunto de datos. En esta exploración de nulos se observo que la columna barrio, contiene datos nulos en aproximadamente 10% de sus registros.

```
#Buscamos nulos

MLDF.select([count(when(isnan(c)|col(c).isNull(),c)).alias(c) for c in MLDF.columns]).display()
```

Grafica 5 – Código de detección de registros nulos en las columnas del conjunto de datos.

Tras identificar 134 valores nulos en barrio. Al no ser una variable necesaria en el análisis se optó por eliminarla, al igual que las columnas cultura_pueblo_o_rasgo y

nombre_institución_nacimiento, como se observa en la Grafica 6 - Eliminación de columnas. Una vez finalizado, posteriormente se eliminaron los demás valores nulos para tener un conjunto de datos optimo.

```
#Observamos que la columna barrio es la que contiene la mayor cantidad de elementos nulo.
#Debido a que barrio, no representa una variable significativa para el comportamiento de los dat
#Repetimos este procedimiento con variables como cultura_pueblo_o_rasgos_f y nombre_institucion_

MLDF = MLDF.drop("cultura_pueblo_o_rasgos_f")
MLDF = MLDF.drop("nombre_instituci_n_nacimiento")
MLDF = MLDF.drop('barrio')

MLDF.select([count(when(isnan(c)|col(c).isNull(),c)).alias(c) for c in MLDF.columns]).show()
```

Grafica 6 - Eliminación de columnas

- Instalación y eliminación de acentos con Unicode: En la Grafica 7 – Eliminación de caracteres especiales con , se muestra el procedimiento para eliminar los caracteres especiales, lo cual es más fácil que realizar comparaciones entre cadenas de texto. Esto es crucial para tareas como la fusión de conjuntos de datos, búsqueda de coincidencias y agrupamiento de datos además que mejora la manipulación de datos, ya que, muchas operaciones de análisis de datos y machine learning funcionan mejor con datos simplificados. Por ejemplo, al tokenizar texto o realizar análisis de frecuencia de palabras, los caracteres especiales pueden introducir ruido innecesario.

```
from unidecode import unidecode

# Crear una sesión de Spark
spark = SparkSession.builder \
    .appName("Remove Accents") \
    .getOrCreate()

# Convertir el DataFrame de PySpark a un DataFrame de Pandas
pandas_df = MLDF.toPandas()

# Definir una función para eliminar las tildes de una cadena
def remove_accents_str(s):
    return unidecode(s) if isinstance(s, str) else s

# Aplicar la función a cada celda del DataFrame de Pandas
pandas_df = pandas_df.applymap(remove_accents_str)

# Convertir el DataFrame de Pandas de vuelta a un DataFrame de PySpark
MLDF = spark.createDataFrame(pandas_df)

# Mostrar el DataFrame resultante
MLDF.display()
```

Grafica 7 – Eliminación de caracteres especiales con Unicode

- Conversión de tipo de datos: Anteriormente en el proceso de unificación de los conjuntos de datos y tras imprimir el esquema del conjunto de datos definitivos se observo que algunos datos numéricos, estaban en formato de texto, por lo cual, se procedió a convertirlos en formato numérico para mejorar la calidad en posteriores etapas del procesamiento., proceso el cual se evidencia en la Grafica 8 – Conversión de formato textual a numérico.

```
from pyspark.sql.types import IntegerType

# Convertir la columna 'tiempo_de_gestaci_n' a tipo entero
MLDF = MLDF.withColumn("tiempo_de_gestaci_n", MLDF["tiempo_de_gestaci_n"].cast(IntegerType()))

# Convertir la columna 'talla_nacimiento' a tipo entero
MLDF = MLDF.withColumn("talla_nacimiento", MLDF["talla_nacimiento"].cast(IntegerType()))

# Convertir la columna 'n_mero_de_consultas_prenatales' a tipo entero
MLDF = MLDF.withColumn("n_mero_de_consultas_prenatales", MLDF["n_mero_de_consultas_prenatales"].cast(IntegerType()))

# Convertir la columna 'peso_nacimiento' a tipo entero
MLDF = MLDF.withColumn("peso_nacimiento", MLDF["peso_nacimiento"].cast(IntegerType()))

# Crear una nueva columna 'factor_rh_binario' con valores binarios
MLDF = MLDF.withColumn("factor_rh_binario", when(MLDF["factor_rh"] == "positivo", 1).otherwise(0))
```

Grafica 8 – Conversión de formato textual a numérico.

- Convención de datos: En el contexto colombiano, se conoce a la educación media secundaria por diferentes denominaciones, debido a que el conjunto de datos esta conformado por dos municipios que si bien albergan varias similitudes, pertenecen a dos regiones diferentes, por lo cual algunos registros textuales contienen información equivalente en denominaciones distintas, por lo cual se procedió a realizar una convención de dichos registros como se muestra en la Grafica 9 - Convención para la denominación de educación media.

```
from pyspark.sql.functions import when #Importamos when para identificar condiciones al interiror del dataframe de PySpark
#Debido a que los dataframe inciales fueron unidos, algunos niveles de educación equivalentes aparecen bajo otra denominación, por lo cual procedemos a
#Reemplazar "media academica" por "basica secundaria" en la columna nivel_educativo_padre

MLDF = MLDF.withColumn("nivel_educativo_padre",
                       when(MLDF["nivel_educativo_padre"] == "MEDIA ACADEMICA O CLASICA",
                            "BASICA SECUNDARIA")
                       .otherwise(MLDF["nivel_educativo_padre"]))
```

Grafica 9 - Convención para la denominación de educación media

- Filtrado de variables: Como último paso en la etapa de preparación de los datos, se realiza una nueva revisión sobre las columnas existentes al interior del conjunto de datos y se determina que algunas variables no son relevantes para el procesamiento de los datos, ya sea por su naturaleza, o porque ya se encuentran implícitas en otra variable, por lo cual se opta por eliminarlas del mismo, como se observa en la Grafica 10 - Eliminación de variables no relevantes.

```
MLDF = MLDF.drop("fecha_nacimiento")
MLDF = MLDF.drop('departamento_nacimiento')
```

Grafica 10 - Eliminación de variables no relevantes

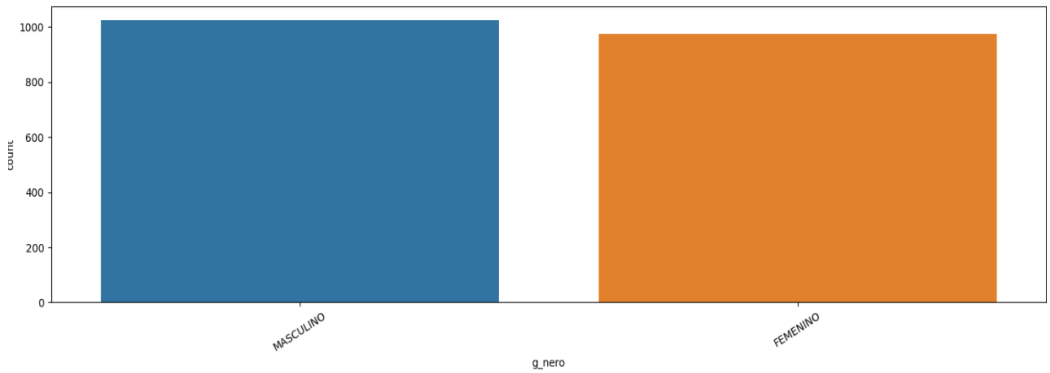
- **Análisis exploratorio de los datos**

  Tras haber culminado exitosamente la preparación de los datos disponibles en el conjunto de datos del proyecto se procedió a realizar una análisis exploratorio del mismo, con el propósito de identificar algunas tendencias, patrones o cualquier otra observación que se pueda desarrollar a través de la visualización de graficas comunes de fácil entendimiento.

  Empezamos por revisar las gráficas producto de las variables almacenadas en el conjunto de datos, a continuación, en la Grafica 11 - Cantidad de datos por municipio presentes en el conjunto de datos, se logra observar la similitud de registros comprendidos en el conjunto de datos resultante de ambos municipios.

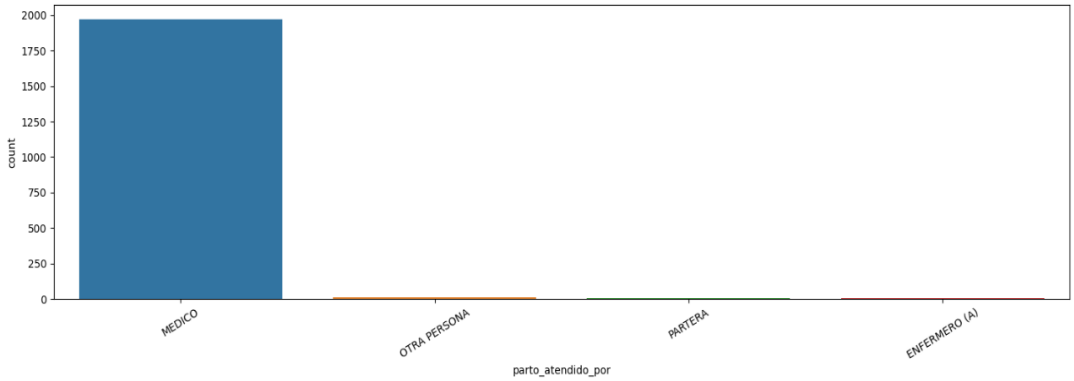
Grafica 11 - Cantidad de datos por municipio presentes en el conjunto de datos

La siguiente grafica corresponde a la distribución de genero en los nacimientos registrados en el conjunto de datos, de igual forma se refleja una similitud bastante alta en la información disponible, como se logra observar en la Grafica 12 - Cantidad de datos por genero.



Grafica 12 - Cantidad de datos por genero

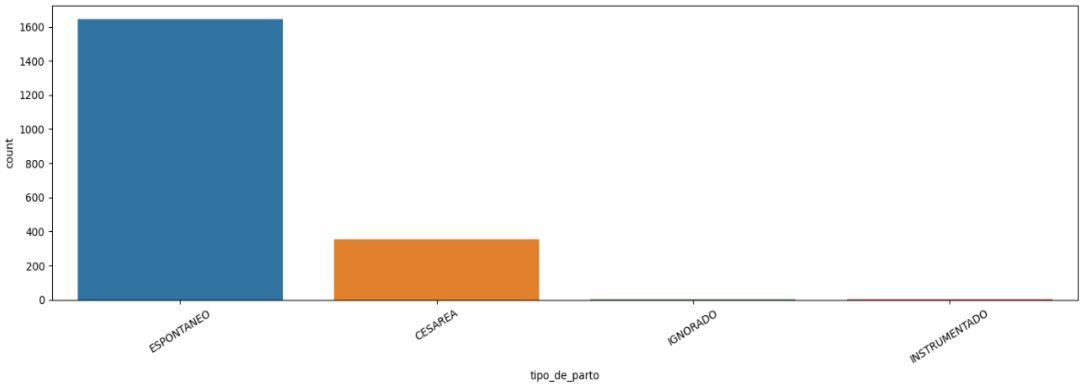
A continuación, en la Grafica 13 - Cantidad de datos por personal de atención en el parto se muestra la profesión que atiende el parto, con una gran tendencia a los partos ser atendidos por médicos, esto es bastante usual, debido a que los conjuntos de datos corresponden a información recolectada en instituciones de salud como hospitales y centros médicos, sin embargo, no refleja la realidad de la totalidad de nacimientos que se producen a diario en la sociedad.



Grafica 13 - Cantidad de datos por personal de atención en el parto

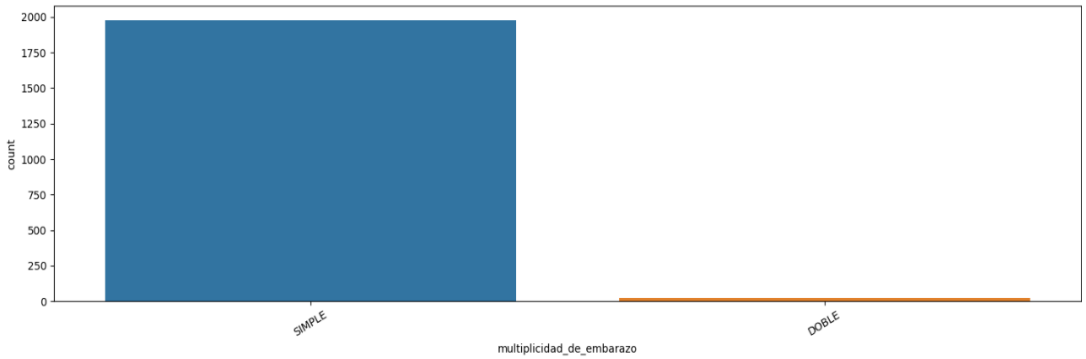
Continuando con el análisis de las gráficas obtenidas por las columnas presentes en el conjunto de datos resultante de la preparación de los datos, observamos en la

Grafica 14 - Cantidad de datos según el tipo de parto, el tipo de parto presente en los registros de nacimientos vivos en los municipios de Acacias y Buga, en esta grafica resaltan como los tipos de parto más comunes son los denominados espontáneos o naturales, sin embargo, también se encuentra en una gran cantidad los partos vía cesárea, mientras que los partos instrumentados no son muy comunes, incluso siendo superados por los registros ignorados, los cuales no registran el respectivo tipo de parto, ya sea por omisión del personal que haya tendido el parto o por omisión de la institución en sus registros.



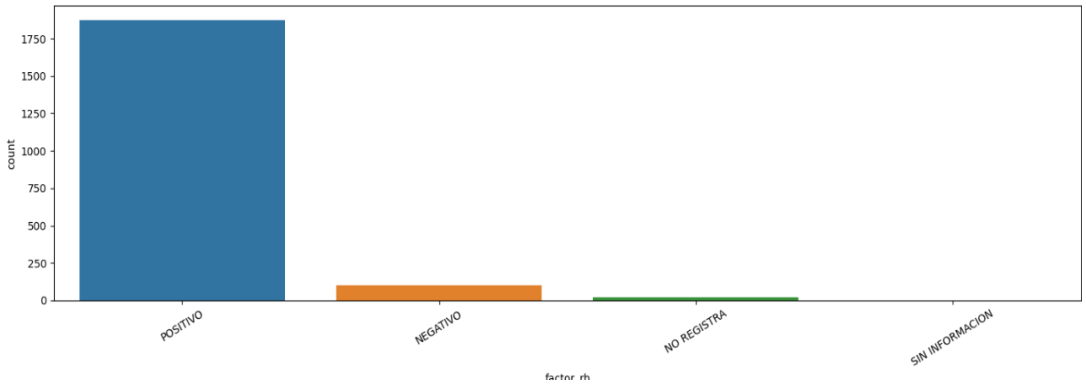
Grafica 14 - Cantidad de datos según el tipo de parto

A continuación, se muestra la Grafica 15 - Cantidad de datos por multiplicidad de embarazo, la cual resulta bastante interesante, ya que es la correspondiente a la variable presente en el conjunto de datos como multiplicidad de embarazo, esta variable es de suma importancia, ya que es una de las cuales participa en las preguntas formuladas para el proyecto en la predicción de los nacimientos a través del uso de algoritmos de machine learning o aprendizaje de maquina en español. En dicha grafica podemos observar la predominancia de los embarazos simples, es decir en los que solo resulta un nacimiento, mientras que se obtienen algunos registros de nacimientos dobles, sin embargo, resulta interesante la falta de un embarazo triple, ya que la totalidad de registros resultantes de la unión de los conjuntos de datos es de alrededor de 8 mil registros, y según la revista médica de la clínica Los Condes, el embarazo triple esta presente en uno de cada 6400 embarazos aproximadamente [1], cabe resaltar que debido al uso de la plataforma DataBricks, el conjunto de datos se limita a 2000 registros a pesar del tamaño de los conjuntos de datos fuente utilizados.



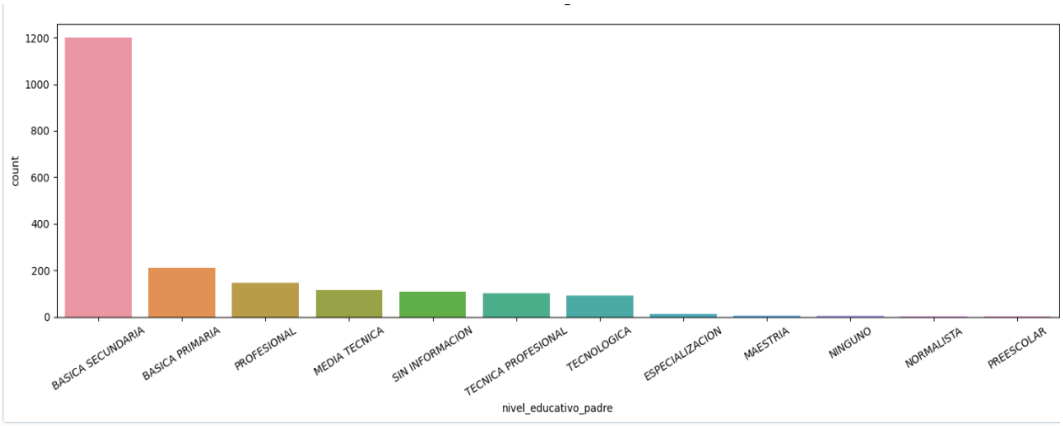
Grafica 15 - Cantidad de datos por multiplicidad de embarazo

Para la Grafica 16 - Cantidad de datos por RH del nacimiento, se muestra la cantidad de nacimientos según el RH, nuevamente existen algunos registros que no cuentan con esta información, y utilizan la denominación No registra o sin información, este tipo de dato es crucial para el desarrollo del bebe, en el aspecto médico, y puede incurrir en graves efectos para su salud posteriormente al no identificar correctamente el RH presente en su sangre, la omisión de estos registros puede ser señal de desatención por parte de la institución o el personal que atiende el parto.



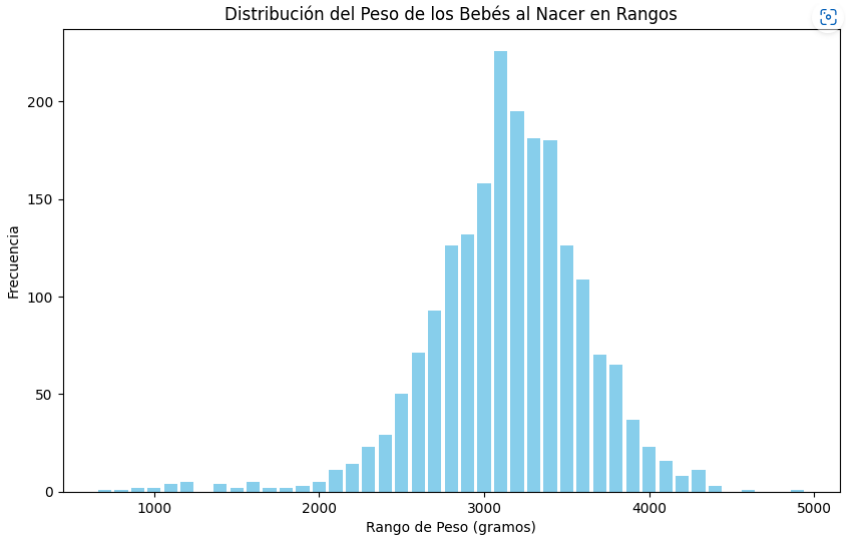
Grafica 16 - Cantidad de datos por RH del nacimiento

Continuando con el recorrido por las graficas obtenidas en el análisis exploratorio de los datos presentes en el conjunto de datos utilizado para el parcial, se encuentra la Grafica 17 - Cantidad de datos por nivel educativo del padre, la cual evidencia una enorme brecha en las personas que tienen acceso a la educación superior en estos municipios, lo cual puede ser una temática de interés a tratar por los gobiernos municipales y departamentales, mas sin embargo no representa mayor relevancia para el actual estudio.



Grafica 17 - Cantidad de datos por nivel educativo del padre

A continuación modificamos la naturaleza de las gráficas por una de distribución, para lograr identificar patrones en los datos según algunas de las variables presentes, por ejemplo, en la Grafica 18 - Distribución de peso de los nacimientos, encontramos que la tendencia en el peso de los nacimientos se muestra hacia un peso normal, lo cual es buen síntoma para la salud del bebe en la etapa posterior al parto, y también encontramos una frecuencia en algunos pesos fuera de los índices normales[2]. Si bien, esta frecuencia es mínima, está presente en algunos nacimientos, lo cual indica puede llegar a indicar un alto riesgo en la salud del bebe nuevamente en la etapa posterior al parto.



Grafica 18 - Distribución de peso de los nacimientos

A raíz de esta observación de los pesos, se comprendido la necesidad de una nueva variable derivada, denominada alto riesgo, la cual toma en cuenta los pesos de los

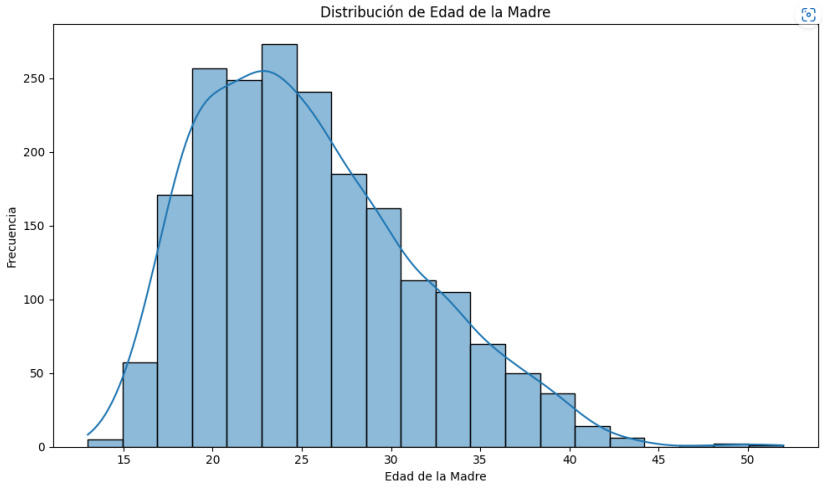
nacimientos por fuera del rango de 2200 gramos y los 4000 gramos, para caracterizar a un nacimiento de alto riesgo, de igual forma, también se puede llegar a considerar un embarazo múltiple como un embarazo de alto riesgo [3], por lo cual, como se observa en la Grafica 19 - Creación de variable derivada alto riesgose procedió a su creación al interior del conjunto de datos, para brindar una relevancia al estudio.

```
from pyspark.sql.functions import when

# Agregar una nueva columna 'Alto_Riesgo'
MLDF = MLDF.withColumn('Alto_Riesgo', when((col('peso_nacimiento') > 4000) | (col('peso_nacimiento') < 2200) | (col('multiplicidad_de_embarazo') == 'DOBLE'), 1).otherwise(0))
```

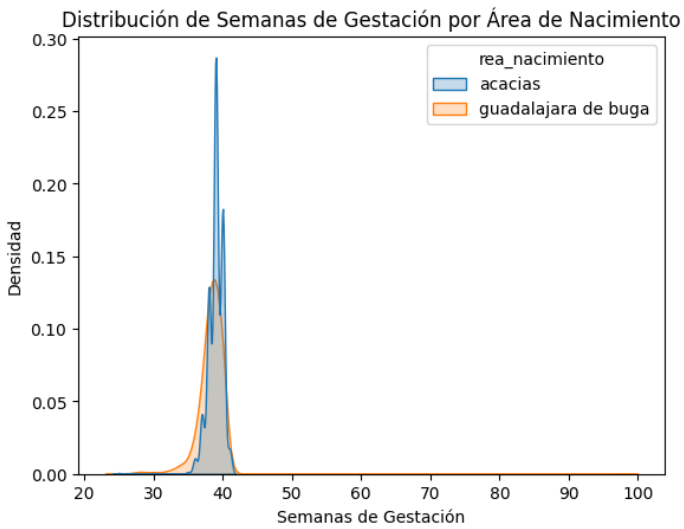
Grafica 19 - Creación de variable derivada alto riesgo

Al proceder con la siguiente variable disponible en el conjunto de datos para el desarrollo del parcial, encontramos la Grafica 20 - Distribución de edad de la madre, la cual muestra una tendencia de madres a temprana edad en un rango entre los 20 a los 30 años de edad, con algunas distribuciones fuera de este rango, obedeciendo a una naturaleza de sesgo a la izquierda en la grafica de los registros presentes en el conjunto de datos.



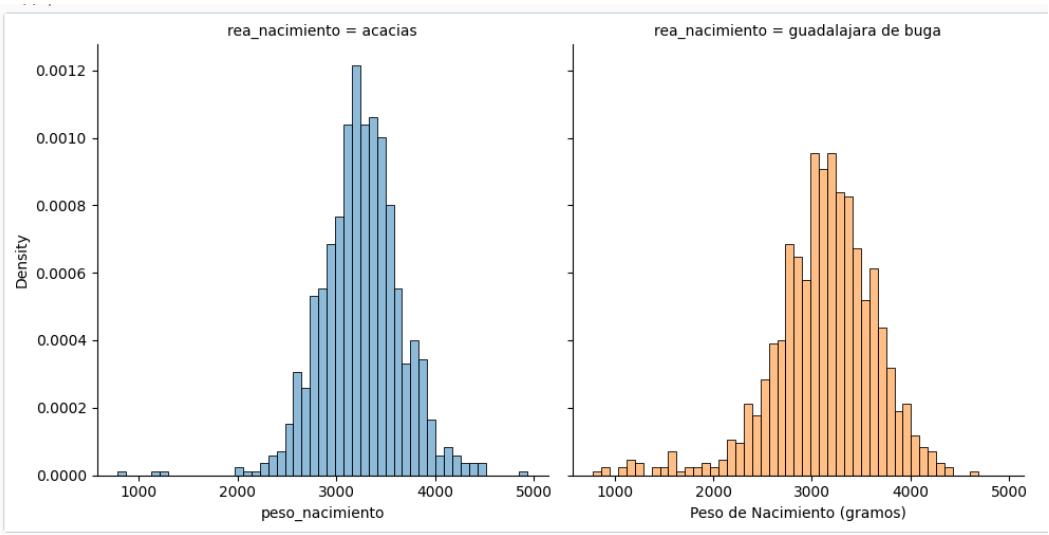
Grafica 20 - Distribución de edad de la madre

Ahora en la Grafica 21 - Distribución de las semanas de gestación según el municipio, se muestra una comparativa, la cual se enfoca en la distribución de las semanas de gestación según el municipio, allí encontramos que la densidad de las semanas de gestación es mucho mas cercana al promedio en el municipio de Acacias, mientras que las semanas de gestación en el municipio de Buga se encuentran un poco mas distribuidas en valores cercanos al promedio pero no tienden a cumplir el mismo.



Grafica 21 - Distribución de las semanas de gestación según el municipio

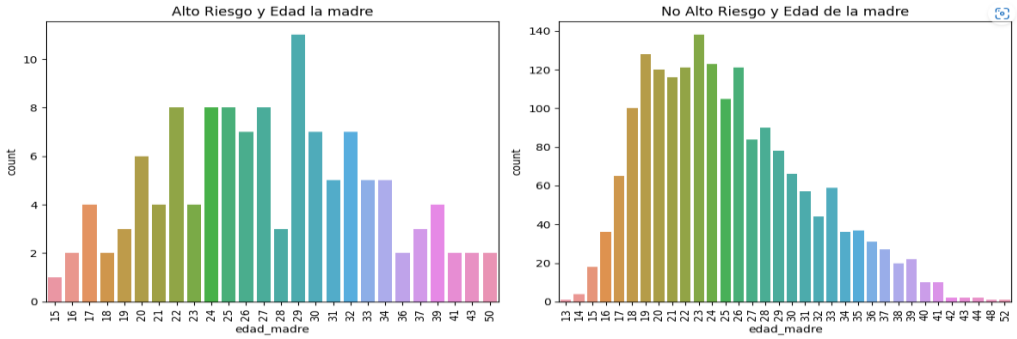
Continuando con las gráficas comparativas, en la Grafica 22 - Comparativa en la distribución del peso de los nacimientos según el municipio, esta vez observamos como en ambos municipios existe una tendencia hacia el cumplimiento del peso promedio en un nacimiento con una mayor densidad en el municipio de Acacias, sin embargo se evidencia de igual forma una mayor densidad en valores considerados de alto riesgo para el municipio de Buga, lo cual puede ser una problemática interesante a tratar por parte del gobierno municipal de Buga.



Grafica 22 - Comparativa en la distribución del peso de los nacimientos según el municipio

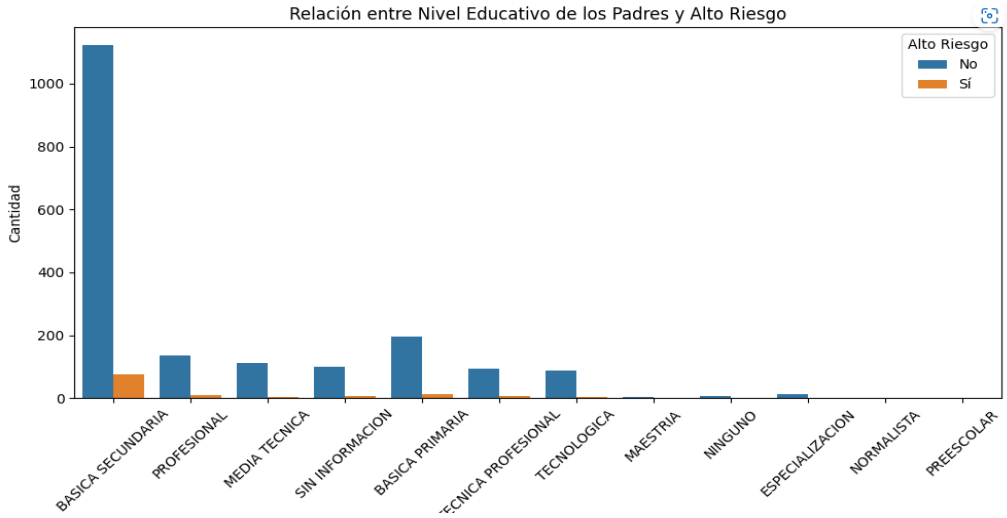
Nuevamente en el análisis exploratorio resulta conveniente consultar una gráfica comparativa, esta vez no entre municipios, sino en la variable derivada alto riesgo,

como se muestra en la Grafica 23 - Comparación de la distribución del alto riesgo del nacimiento según la edad de la madre, donde se analiza el riesgo del nacimiento dada la edad de la madre, esta grafica resulta interesante, ya que si bien se observa que la tendencia de no riesgo esta presente en madres de una edad entre los 17 y 26 años de edad, esto obedece más a la cantidad de datos disponibles para embarazos en madres de dicha edad, ya que en el índice de alto riesgo, se encuentra una distribución bastante equilibrada con un pico en la edad de 29 años, la cual, no es una edad significativamente alta para ser determinada como factor de riesgo.



Grafica 23 - Comparación de la distribución del alto riesgo del nacimiento según la edad de la madre

Y para finalizar el análisis exploratorio de los datos, encontramos la Grafica 24 - Cantidad de datos según la relación entre el riesgo del nacimiento y el nivel educativo del padre, la cual no muestra una relación directa, mas bien obedece a la cantidad de registros de existentes dentro del conjunto de datos con una característica y es la falta de acceso a la educación superior por parte de los hombres en ambos municipios.



Grafica 24 - Cantidad de datos según la relación entre el riesgo del nacimiento y el nivel educativo del padre

b. Problema

- **Descripción del problema**

Las dos preguntas seleccionadas para trabajar corresponden a los siguientes ámbitos:

- Predicción de circunstancias según características del recién nacido, en la pregunta: ¿Se puede predecir el tipo de nacimiento de un recién nacido de los municipios de Buga y Acacias a través de la utilización de un modelo de clasificación? Ya que, a partir de la formulación de un algoritmo de aprendizaje de máquina, de naturaleza de clasificación, podemos obtener el tipo de nacimiento, esperando una exactitud bastante considerable, lo cual puede apoyar la toma de soluciones en algunos aspectos del tratamiento a los nacimientos de los municipios.
- El segundo ámbito es la identificación de factores de riesgo, en este ámbito es donde se aterriza la pregunta: ¿Se puede predecir a través de un modelo de clasificación cuando un recién nacido hace parte de un nacimiento múltiple? Esta pregunta resulta en un interés particular, ya que podría ayudar a determinar, en algunos casos hipotéticos la multiplicidad de un embarazo, lo cual contribuye a tener atención especial en estos casos y reducir en gran proporción el riesgo de los embarazos o nacimientos. Esta predicción es igualmente hecha a partir de un algoritmo de aprendizaje de maquina de clasificación, por lo cual resulta útil, ya que en ambos escenarios se pueden utilizar los mismos algoritmos lo cual supone un ahorro de recursos, debido a que los entrenamientos pueden obedecer a circunstancias similares, además de que el requerimiento de investigación se reduce un poco.

El planteamiento de estos ámbitos se aterriza en congruencia con las necesidades que atraviesan los habitantes de estos municipios, ya que gran parte de la población presente en estas cabeceras municipales pertenece a los rangos de edad en los cuales la paternidad es una opción presente y este tipo de estudios involucrando tecnologías de vanguardia como el aprendizaje de maquina y la inteligencia artificial elevan el nivel de conocimiento respecto al tema tanto de los habitantes del municipio como de los afectados por las políticas que puedan llegar a tomarse gracias a los resultados que arroje el estudio. También este tipo de avances en la investigación de factores reales donde se aplican conceptos académicos a la realidad de una sociedad sirven como estandarte para la implementación de iniciativas no solo en el municipio, sino que incentiva la implementación en las municipalidades aledañas

además, e incluso a nivel nacional, lo cual puede conllevar a una creciente necesidad de profesionales en áreas tecnológicas, como Ciencia de Datos e Ingeniería de Sistemas, profesiones las cuales presentan grandes campos de acción y oportunidades, para estos pueblos que como se menciono anteriormente en el documento, tienen índices preocupantes en cuanto a la posibilidad de sus habitantes en tener acceso a la educación superior.

- **Problemas típicos en datos**

El contenido de esta sección presenta mayor relación con la etapa de preparación de los datos descrita anteriormente, por lo cual es recomendable llevar una relación de los contenidos presentados en ambas secciones, para efectos como la evasión de duplicidad en la información, se presentará una breve descripción de los problemas tratados, los cuales fueron explicados a mayor detalle anteriormente en la sección mencionada, si el lector desea profundizar en alguno de los problemas puede retroceder a dicha sección o consultar el cuaderno adjunto a este documento, el cual también se encuentra disponible en línea, en las referencias presentes en este documento.

Durante la realización del parcial, nos encontramos con bastantes situaciones particulares del abordaje de datos, los cuales, si bien representan una carga de trabajo considerable, no requieren de mayor conocimiento o interpretación para su resolución. Anteriormente en la sección de preparación de los datos y durante la exploración de estos, se evidencio el tratamiento de algunos de estos problemas, entre los cuales destacan:

- Datos faltantes: En la preparación de datos se mostro como en algunas columnas se encontraban datos faltantes, sin embargo en la mayoría de columnas se limitaban a uno o dos registros, por lo cual se considero que su eliminación no influye en gran proporción los resultados que se pudieran presentar, por lo cual, se trato este problema a través de la eliminación de registros con datos nulos, adicionalmente a esto, la columna con mayor porcentaje de datos nulos o faltante, alrededor de un 10%, fue barrio, columna la cual, no conllevaba mayor relevancia para el estudio adelantado por lo cual fue eliminada del conjunto de datos.
- Datos inconsistentes: Este problema, también fue evidenciado durante la preparación de los datos, en la denominación del mismo nivel educativo de distintas maneras debido a la realidad de regiones del país, por lo cual, se opto por unificar estas

denominaciones en una sola convención de variables para no incurrir en disminución de los registros.

- Creación de nuevas variables: Para la creación de una nueva variable, durante la exploración de datos se encontró la necesidad de una variable que identificara el nacimiento según su riesgo, este tipo de variable corresponde a una clasificación binaria, la cual fue lograda a través de la consulta de algunas fuentes confiables en temáticas de salud, las cuales proporcionaron parámetros claros para la adopción de la nueva variable. Esta variable, describe que si un nacimiento se encuentra por fuera de los rangos aceptables de peso del recién nacido, o si él nacimiento hace parte de un embarazo múltiple se puede considerar como un nacimiento de alto riesgo, lo cual no solo indica que se corra riesgo durante el embarazo, sino que además el recién nacido requiere de una mayor gama de cuidados en su periodo posterior al parto, en la Grafica 19 - Creación de variable derivada alto riesgo, se encuentra el código requerido para la creación de dicha variable.
- Fusión de datos: Este problema detallado en la sección de preparación de los datos, fue ampliamente resuelto por el proceso previo de estandarización en las denominaciones de las variables o columnas de los conjuntos de datos, previos a la unión además de la similitud de los conjuntos de datos antes de cualquier tipo de preparación. Se resolvió además del procesamiento previo, que las columnas que no fueran mutuas entre los conjuntos de datos, deberían quedar por fuera del estudio, ya que esto podría contribuir a un sesgo en la generación de resultados, puesto a que la presencia de datos en solo uno de los municipios, implicaría ruido en la interpretación del otro municipio, además que la opción a convenir en dicha situación seria la imputación de valores, lo cual es ideal para pocos registros, pero no lo es para todo un conjunto de datos, debido a que el valor a imputar incluso puede alejarse demasiado de la realidad del municipio en sí.

c. Implementación de Machine Learning y resultados

La implementación de machine learning o aprendizaje de maquina en español, es fundamental para la realización de este parcial, debido a la naturaleza de las preguntas planteadas para su resolución.

Y la consecución de esta etapa se llevo gracias a la realización de las etapas anteriormente descritas en el presente documento, ya que la calidad de los datos, representa la mayor parte del trabajo en la implementación de este tipo de tecnología, las cuales si bien puede llegar a implementarse en escenarios donde la calidad de datos no es la ideal, su efectividad podría ser puesta en tela de juicio.

Para la implementación de esta fase se opto por los siguientes algoritmos:

- Regresión logística: La regresión logística es una técnica de análisis de datos que utiliza las matemáticas para encontrar las relaciones entre dos factores de datos. Luego, utiliza esta relación para predecir el valor de uno de esos factores basándose en el otro. Normalmente, la predicción tiene un número finito de resultados, como un sí o un no. La regresión logística es una técnica importante en el campo de la inteligencia artificial y el machine learning (AI/ML). Los modelos de ML son programas de software que puede entrenar para realizar tareas complejas de procesamiento de datos sin intervención humana. Los modelos de ML creados mediante regresión logística ayudan a las organizaciones a obtener información procesable a partir de sus datos empresariales. Pueden usar esta información para el análisis predictivo a fin de reducir los costos operativos, aumentar la eficiencia y escalar más rápido [4].
- Arboles de decisión: Es un algoritmo de aprendizaje supervisado no paramétrico, que se utiliza tanto para tareas de clasificación como de regresión. Tiene una estructura de árbol jerárquica, que consta de un nodo raíz, ramas, nodos internos y nodos hoja. un árbol de decisión comienza con un nodo raíz, que no tiene ramas entrantes. Las ramas salientes del nodo raíz alimentan los nodos internos, también conocidos como nodos de decisión. En función de las características disponibles, ambos tipos de nodos realizan evaluaciones para formar subconjuntos homogéneos, que se indican mediante nodos hoja o nodos terminales. Los nodos hoja representan todos los resultados posibles dentro del conjunto de datos. [5].

Para la implementación de ambos algoritmos se utilizo la biblioteca de PySpark disponible en Python, en el paquete pyspark.ml.classification, esta biblioteca permite el uso de estos modelos de aprendizaje de maquina de forma fácil, a través de la implementación de un conjunto de datos y un vector de características.

- **Entrenamiento de los modelos**

A continuación, se muestran algunos resultados obtenidos durante la etapa de entrenamiento de cada uno de los modelos, cabe destacar que para el entrenamiento de ambos modelos se utilizaron en ambos contextos la totalidad de variables disponibles al interior del conjunto de datos, salvo la variable objetivo, en cada una de las preguntas:

Algoritmo utilizado	Tiempo de entrenamiento
Arboles de decisión	1.42 minutos
Regresión logística	8.76 minutos

Tabla 1 - Rendimiento de entrenamiento de algoritmos en DataBricks – Pregunta 1

El comportamiento observado en la Tabla 1 - Rendimiento de entrenamiento de algoritmos en DataBricks – Pregunta 1, muestra el rendimiento de cada uno de los algoritmos durante su entrenamiento en el entorno de ejecución de DataBricks, donde la infraestructura nos limita a la capa gratuita, la cual presenta un procesador de dos núcleos, pero del cual se desconoce tanto la velocidad de dichos núcleos, como las demás características del procesador y aproximadamente 16 GB de memoria RAM, nuevamente se volvió a realizar el entrenamiento cambiando a variable objetivo para la pregunta dos, y los resultados se reflejan en la Tabla 2 - Rendimiento de entrenamiento de algoritmos en DataBricks - Pregunta 2 a continuación:

Algoritmo utilizado	Tiempo de entrenamiento
Arboles de decisión	1.38 minutos
Regresión logística	8.03 minutos

Tabla 2 - Rendimiento de entrenamiento de algoritmos en DataBricks - Pregunta 2

Debido a la naturaleza de estos resultados, se resolvió intentar de nuevo el entrenamiento, esta vez en otras condiciones de infraestructura, en un dispositivo de computo personal, el cual dispone de los siguientes recursos: Procesador Intel Core i5 – 8300H @2.30 Ghz, 8 GB de memoria RAM y Procesador Grafico Nvidia GTX 1050 4GB.

Si bien este dispositivo, ya obedece a más de media década de antigüedad, y se ve en escenario de inferioridad en cuanto a un recurso tan importante como la RAM, la

disponibilidad de ejecución en un ámbito nativo, puede llegar a variar significativamente el rendimiento, como se refleja a continuación:

Algoritmo utilizado	Tiempo de entrenamiento
Arboles de decisión	18 segundos
Regresión logística	1.42 minutos

Tabla 3 - Rendimiento de entrenamiento de algoritmos en PC – Pregunta 1

Como se refleja en la Tabla 3 - Rendimiento de entrenamiento de algoritmos en PC, el rendimiento evidencia una mejora sustancial, esto se debe a la disponibilidad de un procesador de gráficos como el presente en el PC, el cual optimiza este tipo de tecnologías en especial los procesadores de la marca, ya que cuentan con tecnologías especializadas para aumentar el rendimiento de los algoritmos, es por esto que este tipo de procesadores son los predilectos a la hora de entrenar algoritmos de una escala considerable.

Algoritmo utilizado	Tiempo de entrenamiento
Arboles de decisión	18 segundos
Regresión logística	1.42 minutos

Tabla 4 - Rendimiento de entrenamiento de algoritmos en PC - Pregunta 2

Se repite el procedimiento para la pregunta 2, con resultados similares como se observa en la Tabla 4 - Rendimiento de entrenamiento de algoritmos en PC - Pregunta 2

- **Resultados finales**

Una vez realizado el entrenamiento se procede a conocer los resultados finales obtenidos en la etapa de pruebas de cada uno de los algoritmos, esta etapa fue reproducida solo en el entorno de DataBricks, ya que en esta plataforma se puede comprobar la disponibilidad de los resultados al alcance de cualquier interesado, mientras que en un ambiente local requiere de cierta configuración lo cual no es ideal para el ejercicio. La Tabla 5 - Rendimiento de algoritmos pregunta 1, contiene los resultados finales de cada uno de los algoritmos correspondientes a la pregunta 1.

Métrica	Arboles de decisión	Regresión logística
Exactitud	0.8272	0.8622
Precisión	0.8076	0.8480
Exhaustividad	0.8272	0.8622

Tabla 5 - Rendimiento de algoritmos pregunta 1

A continuación, se presentan los resultados obtenidos en el escenario de pruebas de cada algoritmo haciendo el respectivo cambio de variable objetivo para la pregunta 2.

Métrica	Arboles de decisión	Regresión logística
Exactitud	0.8272	0.8622
Precisión	0.8076	0.8480
Exhaustividad	0.8272	0.8622

Tabla 6 - Rendimiento de algoritmos pregunta 2

La exactitud mide la proporción de instancias correctamente clasificadas (TP + TN) sobre el total de instancias (TP + TN + FP + FN). Indica la fracción de predicciones correctas en el conjunto de datos completo. Es útil cuando las clases están balanceadas.

La precisión mide la proporción de verdaderos positivos (TP) sobre el total de instancias clasificadas como positivas (TP + FP). Indica qué tan bien el modelo evita clasificar incorrectamente una instancia negativa como positiva. Es útil cuando el costo de los falsos positivos es alto.

La precisión mide la proporción de verdaderos positivos (TP) sobre el total de instancias clasificadas como positivas (TP + FP). Indica qué tan bien el modelo evita clasificar incorrectamente una instancia negativa como positiva. Es útil cuando el costo de los falsos positivos es alto.

Cabe resaltar que cada métrica tiene su propia utilidad dependiendo del contexto del problema y del costo asociado a los errores de clasificación. En problemas donde el costo de los falsos negativos es alto (como diagnósticos médicos), la exhaustividad puede ser más importante. En problemas donde el costo de los falsos positivos es alto (como detección de fraudes), la precisión puede ser más relevante. La exactitud es una métrica más general y puede ser menos informativa en casos de clases desbalanceadas.

- **Respuestas**

¿Se puede predecir el tipo de nacimiento de un recien nacido de los municipios de Buga y Acacias a través de la utilización de un modelo de clasificación?

Ciertamente se puede, para el contexto del parcial, se puede optar fácilmente por el algoritmo de arboles de decisión que si bien no presenta las mejores métricas, el ahrro en un recurso tan importante si no es que el mas importante como lo es el tiempo, justifica una reducción en las métricas de rendimiento obtenidas en el apartado de pruebas, además de que para un ambiente de ejecución como DatBricks puede ser pertinente reducir al máximo este tipo de procesos ya que pueden exceder las capacidades del entorno.

¿Se puede predecir a través de un modelo de clasificación cuando un recién nacido hace parte de un nacimiento múltiple?

Efectivamente se puede, nuevamente para el contexto del parcial, lo ideal es el algoritmo de arboles de decisión por las razones previamente mencionadas, además la implementación de esta respuesta en cualquier institución de los dos municipios estudiados puede significar una gran ventaja competitiva de dicha institución frente a las demás que ejerzan en la zona.

- **Conclusiones, observaciones y recomendaciones**
- A partir de los datos obtenidos, no hay una relación directa entre la caracterización del riesgo en el embarazo y la edad de la madre
- Podrían explorarse más modelos de ML para proporcionar una respuesta mucho más precisa.
- Si bien la calidad de los datos antes de su procesamiento es bastante buena, tras este, los datos dan lugar a muchos más tipos de análisis.
- Las municipalidades de Acacias y Buga pueden trabajar en base a los resultados de este estudio en mejorar las condiciones de acceso a la educación superior de la población varonil.
- En general, las condiciones de los nacimientos en Acacias presentan una mejor expectativa de la salud de los bebes, en comparación a los de Buga.

Referencias

[1] “Embarazo gemelar”. Elsevier | Un negocio de análisis de información. Accedido el 20 de mayo de 2024. [En línea]. Disponible: https://www.elsevier.es/es-revista-revista-medica-clinica-las-condes-202-articulo-embarazo-gemelar-S0716864014706455#:~:text=La%20frecuencia%20clásicamente%20se%20describe,cada%201.000%20embarazos%20(1).

[2] S. A. Gutiérrez. “¿Cuál es el peso adecuado del bebé en el momento de nacer?” Reproducción Asistida ORG. Accedido el 20 de mayo de 2024. [En línea]. Disponible: https://www.reproduccionasistida.org/pesos-en-el-nacimiento/#peso-del-bebe-al-nacer

[3] “Embarazo múltiple: Gemelos o más bebés | Cigna”. Cigna Healthcare | Health Insurance, Dental Plans & Medicare. Accedido el 20 de mayo de 2024. [En línea]. Disponible: https://www.cigna.com/es-us/knowledge-center/hw/temas-de-salud/embarazo-mltiple-hw236272#:~:text=Cualquier%20embarazo%20tiene%20riesgos.,Presión%20arterial%20alta%20y%20preeclampsia%20.

[4] “¿Qué es la regresión logística? - Explicación del modelo de regresión logística - AWS”. Amazon Web Services, Inc. Accedido el 20 de mayo de 2024. [En línea]. Disponible: https://aws.amazon.com/es/what-is/logistic-regression/

[5] “¿Qué es un árbol de decisión? | IBM”. IBM in Deutschland, Österreich und der Schweiz. Accedido el 20 de mayo de 2024. [En línea]. Disponible: https://www.ibm.com/es-es/topics/decision-trees

[6] “Nacidos vivos en el Municipio de Acacías | Datos Abiertos Colombia”. la plataforma de datos abiertos del gobierno colombiano. Accedido el 20 de mayo de 2024. [En línea]. Disponible: https://www.datos.gov.co/Salud-y-Protecci-n-Social/Nacidos-vivos-en-el-Municipio-de-Acac-as/7ifb-wrqs/data?no_mobile=true

[7] “Guadalajara de Buga : Nacidos Vivos 2021”. Datos Abiertos Colombia | Datos Abiertos Colombia. Accedido el 20 de mayo de 2024. [En línea]. Disponible: https://www.datos.gov.co/Salud-y-Protecci-n-Social/Guadalajara-de-Buga-Nacidos-Vivos-2021/dc6m-g67k/about_data